1977 WL 806
United States District Court, N.D. Alabama, Southern Division.

Ensley Branch of the N.A.A.C.P., Plaintiff
v.
George Seibels et al., Defendants.
John W. Martin et al., Plaintiffs
v.
City of Birmingham et al., Defendants.
United States of America, Plaintiff
v.
Jefferson County et al., Defendants.
Lucy Walker et al., Plaintiffs
v.
Jefferson County Home et al., Defendants.

Civil Action No. CA 74-2-12-S | Civil Action No. CA 74-Z-17-S | Civil Action No. CA 75-P-0666-S | Civil Action No. CA 76-M-2047-S | January 10, 1977

Opinion

POINTER, D.J.

Memorandum of Opinion

***1** Since 1945 the Personnel Board of Jefferson County has been charged under state law with the duty of periodically administering examinations to "fairly test the relative capacity and fitness" of applicants for positions with local governmental agencies.[1] 1940 Ala. Code Appx. §§ 645, *et seq.* (Recomp. 1958). Those who pass are ranked on an eligibility list in the order of their exam scores.[2] As vacancies occur, the three persons then at the top of the list are certified to the employing agency for final selection, the appointments being probationary in nature for the first twelve months.[3] An applicant's name may be removed from the eligibility list after having three times been certified and refused employment.

This litigation challenges the employment practices of the governmental agencies as discriminatory on the basis of race, color, and sex, and includes an attack upon the examinations administered by the Personnel Board. Presently at issue, following a trial held December 20-22, 1976, are the tests currently used to screen applicants for positions as police officers, deputy sheriffs,[4] and firefighters.[5]

An attack upon the police and firefighters exams is certainly understandable when one considers that, although the relevant labor pool is over 25% black, yet on June 30, 1976, only 56 (or 6.5%) of the 860 police officers were black and only 9 (or 1.4%) of the 630 firefighters were black. These statistics may, however, be misleading for purposes of this lawsuit because they include the historical results of hiring practices employed long before passage of the Equal Employment Opportunity Act of 1972 or, indeed, before utilization of the tests under scrutiny at this time.

The principal focus should rather be upon the events of more recent years, with particular attention upon practices subsequent to March 24, 1972, when Title VII of the Civil Rights Act of 1964 was made applicable to the Personnel Board and the governmental agencies which it serves. Likewise, information as to the general labor pool in the area is of only marginal importance when one has, as we do, extensive data as to actual applicants for positions and there is no evidence that minority applications have been depressed by prior employment practices.

Adoption of the Current Tests

***2** In late 1965, following an independent study as to why no blacks were then employed as police officers in the City of Birmingham, the Personnel Board decided to replace its police and firefighter exams with tests developed by the Public Personnel Association, now known as the International Personnel Management Association. IPMA tests were being widely used in other parts of the country and were considered by the Board as superior to other tests then available. The change was part of a multi-faceted program intended to increase black participation in governmental positions. (See Appendix C to X-342). Policeman Test 10-C and Firefighter Test 20-B have been in use since August 18, 1967, and October 23, 1968, respectively, as the screening examinations for these positions under the state-mandated selection procedure,[6] although at times other tests have been administered for experimental purposes or for validation studies. Since April 10, 1974, a modified scoring key (based upon only 80 of the 120 test items) has been employed in grading the 10-C test for purposes of the eligibility list. This modification was made at the recommendation of qualified independent consultants who, after study, concluded that the scoring change would increase validity of the test for black applicants.

Intent

It is clear that the Personnel Board, in performing its functions as an employment agency for the various local governments, has not intentionally discriminated against blacks. Indeed, at least since 1965, the Board has not only sought to provide non-discriminatory opportunities for black applicants, but also attempted, within the limits of its statutory duties, to rectify the racial imbalances in local government employment. Some mention of these aims and efforts is appropriate.

It was the Board's hope that adoption of the tests now in issue would benefit black applicants, while nevertheless providing a fair "test of the relative capacity and fitness" of all applicants, as required by state law. Immediately, a study was undertaken to ascertain whether the IPMA policeman test, although a paper and pencil test, would correlate positively and significantly with a widely used non-verbal performance test of general intelligence, the Revised Beta Examination–and it did. As successful applicants were employed by the city of Birmingham, were trained at the police academy, and entered performance of their duties, information was incorporated into the Board's on-going validation studies–which, while lacking sufficient blacks in the sample (only 6 in the 10-C sample) to permit full analysis, were considered by the Board as justifying further usage of the 10-C.[7] These studies are presented by the Board not as satisfying the requirements of the EEOC or Department of Justice guidelines on tests, but rather as indicating its efforts to see that its examinations were fair predictors of job performance even at a time when it was not subject to the provisions of Title VII. As already noted, when, in 1974, it was advised by independent consultants that a modification of the scoring of the 10-C exam would improve the validity for black applicants, it immediately put that change into effect.

***3** Since 1965 the Board, with the cooperation of local civic groups and some of the employing agencies, has been actively engaged in recruitment efforts to attract black applicants. In 1966 it began assuming the $10.00 medical examination costs for newly hired persons; and in 1967 it was successful in sponsoring legislation to eliminate the $1.50 examination fee previously required and to eliminate the priority previously given applicants who resided within an employing agency's jurisdiction.[8] It has experimented with a lowering of the raw score used to measure a "passing" grade on the exams where it could justify that approach on the basis of "supply" and "demand".

In short, in its selection, administration and use of the 10-C and 20-B tests, there has been no design or intent on the part of the Board to discriminate on the basis of race or color. However, the standard under Title VII of the Civil Rights Act of 1964 is not so limited[9]–rather, if the *operational effect* of test usage is to discriminate against blacks, then it is proscribed unless it is shown to be a "job related" requirement,[10] with a "manifest relation to the employment in question."[11] See U.S.C.A. § 2000e-2(h). In this inquiry the court is to follow the guidelines adopted by the EEOC and, more recently (November 17, 1976), by the Department of Justice (DOJ), absent some "cogent reason."[12] See *Watkins v. Scott Paper Co.*, 530 F.2d 1159 (CA5 1976). Also instructive are the 1974 A.P.A. Standards for Educational & Psychological Tests and the 1975 Principles for the Validation and Use of Personnel Selection Procedures of the A.P.A.'s Division 14.

Adverse Impact

Where the total selection process has an adverse impact upon a substantial racial group in the labor market, the individual components of that process–such as a screening test–are also to be evaluated for adverse impact. DOJ Guidelines § 4b. For

purpose of this two-step analysis, data can be extracted from the evidence pertaining to administrations of the 10-C and 20-B tests which have been used for employment decisions after March 24, 1972:[13]

TABULAR OR GRAPHIC MATERIAL SET AT THIS POINT IS NOT DISPLAYABLE

***4** According to the DOJ Guidelines, § 4b, "A selection rate for any racial * * * group which is less than four-fifths ($^4/_5$) (or eighty percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact * * *. Greater differences in selection rate would not necessarily be regarded as constituting adverse impact where the differences are based on small numbers and are not statistically significant, or where special recruiting or other programs cause the pool of minority * * * candidates to be atypical of the normal pool of applicants from that group."

So far as the total selection process is concerned, one finds from the above data that the hiring rates for blacks (6.6% of the black applicants on 10-C and 3.2% of the black applicants on 20-B) are substantially less than eighty percent of the hiring rates for whites (23.3% and 14.1%, respectively). These greater differences in selection rates cannot be explained on the basis of inadequate numbers and, according to the court's calculations, are statistically significant: the phi coefficient for the 10-C test is .193 and for the 20-B is .121, both of which are significant at rho < .001.

Looking at the data pertinent to the test component of the selection process, one finds again that the pass rates for blacks (48.6% for 10-C and 24.2% for 20-B) are substantially less than eighty percent of the pass rates for whites (90.2% and 82.5%, respectively). And again, according to the court's calculations, these greater differences in pass rates, which are based upon samples of adequate size, are statistically significant: the phi coefficient for the 10-C test is .46 and for the 20-B is .48, both being significant at rho < .001. Also of importance is the fact that, of the blacks who did pass the tests, 85.4% placed in the lower half of the initial eligibility lists for police officers and 89.9% placed in the lower half of the firefighter lists.[14]

Some concern can justifiably be expressed that the special recruiting efforts undertaken by the Personnel Board and other groups to attract black applicants–while commendable as an affirmative action to overcome racial imbalance in the police and firefighter forces–may at the same time have resulted in an atypical pool of blacks taking the test, producing distortion in the test performances of the black applicants. For example, the black applicants may have included many who were not seriously interested or motivated with respect to the jobs in question, thereby affecting their test performances. Absent, however, any hard data to support such an hypothesis or to indicate its magnitude, the court, impressed with the substantial differences in hire rates and pass rates for the two racial groups, must conclude that the overall selection procedures in effect since March 24, 1972, and as a component part thereof the tests used for those purposes, have had an adverse impact on blacks.

Validation Studies

***5** According to EEOC Guidelines § 1607.3, "the use of any test which adversely affects hiring * * * of classes protected by Title VII constitutes discrimination unless (a) the test has been validated and evidences a high degree of utility as hereinafter described * * *." For the purpose of making such validation studies of its many tests, the Board in 1972 contracted with Drs. William E. Farrar and William A. McLaurin, Professors in the Psychology Department of the University of Alabama at Birmingham. Both had experience with personnel selection procedures in public employment systems. Priority, but not exclusive attention, was to be given to the police and firefighter tests, and their work on these tests began in late 1972. Their studies respecting the two tests continued even to the time of trial, with various reports being made in each of the years 1973, 1974, 1975 and 1976. That their work was not complete before trial does not suggest inattention; rather, it is indicative that their studies were intended to be thorough and were directed to numerous tests.[15]

Psychometric Analyses

The 10-C and 20-B tests are paper-and-pencil instruments, each consisting of 120 multiple choice items.[16] The initial concern of Drs. Farrar and McLaurin was directed to the reliability,[17] item difficulty,[18] and item discrimination[19] of the tests. The following findings were made:

TABULAR OR GRAPHIC MATERIAL SET AT THIS POINT IS NOT DISPLAYABLE

***6** Inquiry into reliability is a proper first step, because, while no test is perfectly reliable, a test which is not reliable is not valid for any purpose. The consultants found the reliability coefficients for both tests to be of sufficient magnitude to indicate satisfactory reliability. They did acknowledge that the methods selected for this purpose were essentially measures of internal consistency (and with the KR-20 formula, of content homogeneity), but apparently believed it either not feasible or not necessary to investigate error variance due to time sampling. The court agrees as to reliability and notes that possible lack of stability over time is, in a sense, mitigated by the fact that applicants may take an exam on more than one administration.

Analyses of item difficulty and discrimination have no direct bearing upon the validation studies before the court. However, they do reflect an investigation into possible modification or supplementation of the tests to improve their utility and reduce the extent of adverse impact, which is a recommended procedure.[20] See DOJ Guidelines, § 3c.

Documentation and Methodology

The EEOC Guidelines, at §§ 1607.5(b)(2,3,5) and 1607.6, require that various items of information (*e.g.* copies of tests, manuals, rating forms and instructions and representations of statistical data) be included in the report of the study or otherwise available for inspection. Following the 1974 A.P.A. Standards, a more extensive list of documentation requirements is specified in the DOJ Guidelines at §§ 4a and 13b, involving some twenty-four "essential" items and several other desirable items. The Farrar-McLaurin studies satisfy the EEOC requirements, which were the only ones in effect when their studies were conducted and (so far as then feasible) completed and, indeed, when supplemented by evidence presented immediately before and during trial, they also substantially satisfy the DOJ requirements, which became effective on November 23, 1976.[21]

***7** The studies include presentations of the following statistics:

For the 10-C test:

?? intercorrelation coefficients, r and r_c, for 109 Birmingham police officers (without separation by race) respecting their 10-C scores, police academy scores (school average and course grades) and latest efficiency ratings.

?? means, standard deviations, and *t* tests for difference in means for 10-C scores of 38 black and 101 white Birmingham police officers.

?? means, standard deviations, and r coefficients for 59 Birmingham police officers (without separation by race) respecting their 10-C scores, academy scores (average and courses) and latest efficiency ratings.

?? means, standard deviations, r coefficients, and *t* tests for the following:

?? 20 black and 76 white Birmingham police officers respecting their 10-C scores, academy averages, and latest efficiency ratings.

?? 8 black and 140 white Birmingham police officers respecting their 10-C scores, academy scores (average and courses), latest efficiency ratings (overall and by sub-parts), and experimental ratings weighted average and by components.

?? 49 black and 140 white police officers respecting their 10-C scores and academy scores (average and courses). (Also included are data for analyzing significance of differences in correlation coefficients through z transformations.)

?? 83 Jefferson County deputy sheriffs (without separation by race) respecting their 10-C scores and academy averages.

?? 77 police officers (without separation by race) from other cities served by the Personnel Board respecting their 10-C scores and academy averages.

For the 20-B test: means, standard deviations, and r coefficients for the following:

?? 162 Birmingham firefighters (without separation by race) respecting the 20-B scores, training academy average, and latest efficiency ratings (overall and by sub-parts).

?? 196 Birmingham firefighters (without separation by race) respecting their 20-B scores, academy averages, latest efficiency ratings, and experimental ratings (overall and by components). Statistics are reported separately for short-tenure and long-tenure firefighters, using three years of experience as the point of division.
Statistics found to be significant at rho < .05 and rho < .01 are so identified in the report.

Some comment should be made about selection and composition of the different samples. Each sample contained all the persons for whom, so far as was known at the time by the consultants, the data needed for that study was available. The different studies were, however, conducted over a period of several years as either the need was recognized or the particular inquiry became technically feasible; and during the time intervals the work force had changed. Some of the studies involved concern with additional factors (*e.g.*, performance on the Raven and PAS tests, which have been under consideration for use as supplemental or alternative screening instruments), for whom the data existed only for a limited number of applicants or employees. The result is that a particular sample may contain some, but not necessarily all, of the persons in another sample and may also contain some persons who were not in the other sample. This lack of autonomy or consistency complicates somewhat the process of analysis, but, under the circumstances, is acceptable. There is no hint of contrivance in selection of the samples or of lack of representativeness of the sample subjects.[22]

Criteria

***8** The several criterion-measures (academy grades, efficiency ratings, and experimental ratings) have certain common factors: (1) None appears to be "contaminated" (*i.e.*, affected by knowledge by the rater or scorer of prior score on the 10-C or 20-B). (2) None has been subjected to special statistical scrutiny to detect or control possible bias among raters or graders.[23] (3) None has been subjected to special statistical scrutiny for reliability.[24] (4) Each has been analyzed by the consultants for relevancy (*i.e.*, the extent to which it may be considered as a measure of critical or important work behaviors). Each of the measures has, of course, its own special characteristics and limitations, which will be described separately. It must be emphasized that, in a criterion-related validation study, one is attempting to estimate the extent to which a score on a "predictor" (*e.g.*, 10-C test) can predict job performance (*i.e.*, as a police officer) through evaluating its ability to predict scores or ratings on a "criterion" (*e.g.*, academy average)–and hence the study is subject to any limitations which those same criteria have in either predicting or assessing job performance.

(1) Academy Grades.–Where, as here, new employees are required to complete special training before performing their duties, successful completion of that training may properly be used as a criterion-measure, if, that is, the training is intended to, and does, provide skills or knowledge needed for performance of the job. Based upon the evidence presented, including testimony of the directors of the Birmingham police and fire academies, the court finds that the two schools do serve that purpose and function.

***9** Relative standing or ranking among students who successfully complete such training is not, however, as such, an appropriate criterion.[25] Rather, to be relevant as a criterion, such measures must be shown, empirically or otherwise, to be themselves appropriate predictors of job performance. This, in essence, means a two-step correlation study; and, in a situation where one has data on test scores, academy grades, and measures of job performance for the same group of persons, the more direct inquiry (correlation between test scores and measures of job performance) would be preferred to the two-step approach. Grades on particular courses in the academy must also be analyzed for compatibility with findings respecting grades on other courses.

So far as the evidence indicates, academy grades–provided they are passing scores–have no impact on job opportunities, benefits, etc. If this be the case, then, while helpful in preventing "contamination" during validity studies, academy grades are likely to be influenced by motivational considerations not present in actual job performance. The emphasis in the academies on paper-and-pencil multiple choice items, while providing objectivity, may also reflect a relationship to the paper-and-pencil screening exam not found in job performance. These concerns should cause one to be cautious in making non-empirical judgments about the usefulness of relative academy grades as a criterion-measure.

(2) Efficiency Ratings.–The efficiency ratings given periodically on all employees by their supervisors are direct and, ostensibly, appropriate measures of job performance. Drs. Farrar and McLaurin have, however, acknowledged that these ratings are not trustworthy assessments of the employees' actual performance. In addition to other problems, the ratings must be discussed between the rating supervisor and the employee and can have important consequences for the employee. These

ratings were, it seems, used in the early studies because of their availability, in the anticipation that other measures could be developed and administered in due course.

(3) Experimental Ratings.–By review of existing job descriptions, by interviews to determine "critical incidents" of the jobs, and by technical assistance and consultation with advisory committees consisting of representative incumbents and supervisory personnel, new "experimental" rating forms were developed for use in the Farrar-McLaurin studies. The forms consist of twelve rating categories for each of the two jobs, the categories relating to personality characteristics, job knowledge, and abilities found through the process to be relevant to job performance. Each is rated on a seven-point scale (poor = 1 to outstanding = 7), with 3 being fixed as adequate. For the police form, weights were developed by the advisory committee to indicate relative importance of the categories to overall job performance.

***10** Raters–the employes' supervisors–were given, in person and in writing, standardized instructions for use of the forms. To prevent the "halo" effect, supervisors rated all their subordinates on one category before proceeding to rate them on the next, etc. The raters were told that their evaluations were confidential and would not be used for any purpose other than the evaluation of the tests.

The court is impressed that the experimental rating method so developed represents an appropriate criterion measure for the jobs in question. These jobs are not ones which lend themselves to some objective measure, such as the sales produced by a sales representative. The principal limitations with the ratings so obtained are the lack of evidence as to reliability and the lack of special steps to detect or control possible bias.

Study Findings

Key findings from the Farrar-McLaurin studies are tabulated below. Correlations are shown only where presented in the body or exhibits of their reports. Statistically significant differences in means between two sub-groups (blacks and whites; short-tenure and long-tenure firefighters) are indicated by so designating the lower of the two means.

TABULAR OR GRAPHIC MATERIAL SET AT THIS POINT IS NOT DISPLAYABLE

Fairness and Differential Validity

If members of one racial group generally obtain lower test scores than members of another group and those differences are not reflected in differences in measures of job performance, there is a need, where technically feasible, to investigate for possible unfairness of the test to the first group. See DOJ Guidelines, § 12b(7) (noting that this need increases the greater the severity of the adverse impact on the lowerscoring group). As the tabulation indicates, the Farrar-McLaurin studies do show that blacks as a group have scored lower on the 10-C than have whites. Indeed, in each of their studies where those scores are reported separately for the two racial groups, the differences are significant at rho < .01.

One possibility is that the predictive validity of the test for one racial group is significantly different than for the other group. The inquiry here, as emphasized in A.P.A. Standard E9, is not whether there are differences in the correlation coefficients or whether one coefficient is statistically significant while the other is not. Rather, the proper statistical procedure is to test for significant differences in the coefficients.

In the one study in which a sufficient[26] number of both blacks and whites are involved, Drs. Farrar and McLaurin have performed such an analysis. Using the report of test scores and academy scores for 140 white and 49 black police officers, they tested the correlation coefficients (where the coefficient for either subgroup was significant at rho < .05), after z transformations, for significance of difference. None was significant at rho < .05, and with respect to only one course (accident investigation) was the coefficient significant even at rho < .10.

***11** Failure, however, to reject the hypothesis that the correlation coefficients are the same for both groups is not by itself sufficient to demonstrate fairness. Where, as in the present case, test scores by two groups are used in the same manner for members of both groups, it is on the assumption that in general an individual's test score will appropriately predict his

standing on the criterion whether he is a member of one group or the other. The predictive relationship between the test score and the criterion can be represented by a regression line[27] formula for converting a test score into a predicted criterion score. While regression lines can be calculated separately for the two groups and will almost always be somewhat different, it is important to know whether the slopes or intercepts (or both) of the lines are sufficiently different to call for abandonment of a common line for the two groups. Otherwise, the common regression line (which is the effect of using test scores in the same way for both groups) may systematically underpredict for members of one group (while overpredicting for the other) their criterion score from a particular test score. The method for this inquiry, called analysis of variance, involves use of the *F* distribution tables for statistical significance. If desired, one can determine for what test scores the common regression line should, and should not, be abandoned.

Significantly different regression lines may have the same or similar correlation coefficients. In such a situation comparison of the coefficients will not reveal the inappropriateness of using a common regression line. Thus, with the sample containing 76 white and 20[28] black police officers, comparison of the coefficients respecting 10-C scores (and modified 10-C scores) and either academy averages or efficiency ratings does not, as to any comparison, lead to rejection of the hypothesis of the coefficients being the same. Yet, when the same data are reviewed by analysis of covariance, as the court has done,[29] for the hypothesis that a common regression line fits both whites and blacks, the obtained *F* ratios are 18.38 (10-C and academy average), 3.97 (10-C and efficiency rating), 26.91 (modified 10-C and academy average), and 4.18 (modified 10-C and efficiency rating), each of which (with $n_1 = 1$ and $n_2 = 93$) is significant at rho < .05. It is interesting that in the development of the modified 80-item scoring key for the 10-C, adopted to increase the validity coefficient for blacks, the evidence becomes stronger that a common regression line should not be used for both groups.

***12** In view of the fact that covariance analysis suggests rejection of a common regression line for both racial groups, one is tempted–since the differences between white and black means appear to be far greater on test scores than on the criterion-measures–to conclude that the performance of blacks is being underpredicted by the 10-C. However, if regression lines are computed separately for blacks and whites using the results of any of the studies where their test scores and criterion scores are reported separately, it will be found that the lines cross and that for test scores below that crossing point the criterion scores thereby predicted for blacks are less than for whites for the same test scores. Above that point there would be underprediction for blacks, but the intersections occur at such high test scores (the lowest point from any of the data is at a raw test score of 87) that few blacks would actually be affected. If one looks to see where any overprediction or underprediction is statistically significant, it is found that the only significant range of scores is for lower scores, where blacks are being overpredicted by the 10-C.[30]

The net result is that use of 10-C test scores in the same manner for both blacks and whites does not appear to be underpredicting the performance of blacks at the academy or on efficiency ratings. This analysis does not, of course, deal with the possibility of bias affecting the scores blacks obtain at the academy or on efficiency ratings; it only serves as a foundation for concluding that the 10-C is not to be criticized on the basis of differential validity inquiries. The 20-B cannot be subjected to these inquiries at the present time for lack of sufficient blacks in any study group.

Operational Utility

It is not, however, sufficient that, as here, a test is shown to have a statistically significant relationship to one or more criteria and not to be differentially unfair to the adversely affected racial group. In addition, in the words of the EEOC Guidelines, § 1607.5(c), the relationship between the test and the criterion must have "practical significance" or, in the words of the DOJ Guidelines, § 12b(5), the usage of the test must be evaluated "to assure that it is appropriate for operational use." With different words, the two Guidelines are raising the same concern.

In concluding that the 10-C and 20-B tests are valid screening instruments, Drs. Farrar and McLaurin have emphasized their significant relationship to grades in the training academies, and this relationship cannot be doubted. However, as already indicated, it is the court's conclusion that *relative* standing in the academies, as distinguished from successful completion of academy training, is not an appropriate criterion unless it also be demonstrated that those academy grades are themselves valid predictors of job performance. The studies reflect, however, that for the most part the correlation between academy grades and measures of job performance are not significant and, in the few instances where significant correlations are found, the findings are mixed–some being positive and others being negative. A negative correlation, of course, indicates that the higher the academy grades, the lower the performance ratings tend to be. Although an employer is permitted to select the best person for the job despite resulting impact on a racial group, it is not permitted to engage in such selection procedures merely

to employ the best person for training.

***13** Nor has it here been demonstrated that either test is a valid predictor of successful completion of the required training courses. According to the director of the policy academy, only 11 of the 733 cadets attending the academy since 1962 have failed to complete the training because of inadequate grades–and no data has been presented as to their 10-C test scores. According to the director of the firefighters academy, no student has failed because of poor grades, of course, since historical use of screening tests (whether the present ones or their predecessors) has imposed a restriction of range, the conclusion does not necessarily follow that every applicant could complete the training no matter how low his 10-C or 20-B score. However, it should be noted that on occasion the raw test scores used to determine hiring eligibility have been substantially reduced, without apparent impact on their successful completion of the academies. And–while recognizing that, as noted by Dr. McLaurin, this is not a recommended practice[31]–use of the regression lines developed from any of the studies would predict passing academy averages even for persons scoring zero on the 10-C and 20-B tests.

As earlier discussed, the regular efficiency ratings are not trustworthy criterion-measures of actual job performance. Even if they were, the studies provide inconclusive findings with respect to the 20-B (a significant correlation with a group of 196 firefighters, though only of a magnitude of .20, and an insignificant correlation of .12 with a group of 162 firefighters), and even more dubious results with respect to the 10-C.[32]

The experimental ratings are, as previously indicated, considered by the court as an appropriate criterion measure. The correlation, however, between the 20-B and these ratings is found to be .08, which is, of course, not significant; and, while a significant positive correlation is found with respect to the 103 firefighters having less than 3 years service, a significant negative correlation is found for the 93 having at least 3 years of tenure. Presumably, higher scores on the 20-B (which carried over to higher scores during academy training) resulted in better job performance for the first few years. Had this advantage merely been erased after more time on the job, this would be one matter–but, as stated, the findings actually showed a significant negative correlation for the longer tenured firefighters, suggesting that over time the lower scoring applicants made the better employees. Absent any indication that during the first few years the lower scoring applicants had been inadequate on the job, one is hard pressed to conclude that the higher scoring 20-B applicants are in fact the better persons to hire. Further study might, of course, lead to other interpretations, such as a determination that recent improvements in the training given at the firefighters academy will result in better employees not only initially but also over time–but no such conclusions can be supported on the present evidence.

***14** The correlation between the 10-C and the experimental ratings is, with 148 in the sample, significant at rho < .05, but has a magnitude of only .21. What do these figures mean? To begin with, it should be understood that for a correlation coefficient to be found significant at rho < .05 is equivalent to saying that, if in fact no relationship between the two variables exists for the "population", the obtained results could be expected to occur only once in twenty such samples–and that therefore one can be 95% confident that for the population (of applicants) there is *some* correlation (or relationship) between the two variables. It does not mean that one can be 95% confident that the population coefficient is .21. Indeed, to state the population coefficient with only a 5% chance of error (*i.e.*, pi < .05) requires use of a confidence *interval*: here, with a sample of 148, that the true coefficient lies somewhere between .0504 and .3592. The coefficient obtained from the sample is but an estimate of that true population coefficient.

A second consideration is to look at the test scores in the particular sample in comparison with the scores of all persons in the population. Where, as here, there is a restriction in the range of test scores of those in the sample because of prior use of the test, a statistical technique, called correction for restriction of range, may be appropriate for determining the magnitude of the correlation. This "corrected" coefficient, as reported by Drs. Farrar and McLaurin, is .36. It may be noted that utilization of the correction involves the assumption that the two variables (test scores and experimental ratings) are for the total population "normally distributed;" and, insofar as rating scores are concerned, it is just that–an assumption. Hence, it involves the same type of risk as does the use of a regression formula for values beyond the sample on which based–a technique which, during the trial, provoked Dr. McLaurin's criticism.

In general, other factors remaining the same, the greater the magnitude of the coefficient the more likely it is that the test will be appropriate for use. See DOJ Guidelines, § 12b(5). The importance of the size of the correlation coefficient can perhaps best be understood by reference to certain basic statistical concepts. The square of the correlation coefficient, called the "coefficient of determination," gives the proportion of the variance of the criterion scores which is accountable by reference to variance of scores on the predictor test. Thus, with a correlation coefficient of .21, the study indicates that 4.4% of the variance among experimental ratings is explainable by reference to the variance in test scores, while 95.6% is not. Using the "corrected" coefficient of .36, still only 13% of the variance among experimental ratings could be accounted for by test score variance. By another formula, the correlation coefficient can be converted into a "coefficient of alienation", which gives the

size of the error in attempting to predict experimental rating scores from test scores *relative* to the error that would result from a mere guess, *i.e.*, by not using the test. This calculation, based on a correlation coefficient of .21, reflects that use of the test predicts experimental rating scores with a margin of error that is only 2% smaller than it would be without the test, and, if based on the "corrected" coefficient of .36, indicates that the margin of error is only 6.7% less than what would occur by mere guess.

***15** Anastasi comments, and quite properly so, that evaluation of a test in terms of the error of estimate will for many testing purposes be unrealistically stringent. Anastasi, PSYCHOLOGICAL TESTING, p. 166 (4th Ed. 1976).[33] She notes that even tests with an unusually high validity of .80 would appear to be inefficient if used to predict individuals' relative standing on some criterion, but that most tests are merely used to determine which individuals will exceed a given minimum standard of performance or cutoff point in the criterion. The 10-C, of course, is utilized here both to screen applicants (cutoff scores) and to rank those passing applicants.

While the magnitude of the correlation coefficient is obviously of great importance, there is no minimum coefficient applicable to all employment situations. See DOJ Guidelines § 12b(5). "Under certain circumstances, even validities as low as .20 or .30 may justify inclusion of the test in a selection program." Anastasi, *op. cit.*, p. 166.

Another approach towards evaluation of the relationship found is to investigate the meaning of differences in test scores in relation to the differences in criterion scores thereby predicted. This involves use of the regression formula, which can be calculated from the correlation coefficient and the means and standard deviations of the two variables. Thus, Dr. Roland Ramsey, the plaintiffs' expert witness, questioned the practical value of the 10-C by noting, from the Farrar-McLaurin study involving 140 whites and 49 blacks, that an increase in raw test scores of 40 points produced, under the regression lines given, less than 5 points increase in predicted academy averages.

The common regression line computed for the 148 officers with both test scores and experimental ratings is ?? = 300 + .162x, where x represents a given test score and ?? is the rating predicted thereby. At first glance, this regression line does not appear to be subject to the criticism made by Dr. Ramsey respecting the other study, for it will be seen that, for example, a test score difference of 10 will predict an experimental rating difference of 16. However, it should be understood that the linear regression formula (as well as the criterion mean and criterion standard deviation, although not the correlation coefficient) varies in direct proportion to any factor by which criterion scores in the sample have been multiplied. The Farrar-McLaurin study reports the overall experimental rating as the summation of weighted components which comprise the rating. For example, a sample subject rated as 5 (very good) on each of the 12 rating components would, because of the method, be reported as having a rating score of 499, while another subject identically rated except for scores of 4 (good) on the appearance and dependability components would receive an overall rating of 483, with 698 representing a "perfect" score of 7's on all components.

***16** To prevent potential misinterpretation, it is well to consider the regression line not only in the form in which expressed in the Farrar-McLaurin studies, but also in a form which does not contain the inflation caused by the weighting procedure. This can be done, while still retaining the concept of the components having differing weights, by expressing the weights in a manner in which the average weight is 1. That is, instead of the "communication" component having a weight of 8.78 (as reported in the study), of "problem solving" a weight of 9.44, of "learning" a weight of 8.22, etc., they can be shown as having weights of 1.056, 1.136, and .989, etc., respectively. Then, a summation of weighted component scores for a "perfect" score of 7 on all 12 components would result in 84, identical with a "perfect" score if not weighted. Not only does this method retain the concept of weighing the different components, but the transformation (whether of means, standard deviations, or regression line) can simply be made by dividing the reported results by a constant, here 8.31333. The obtained regression line is ?? = 36.087 + .195x, where x represents the test score and ?? is the predicted rating (using the new method of expressing weights). It will now be seen that, as with the other studies criticized by Dr. Ramsey, a large difference in test scores produces only a small difference in predicted (un-inflated) experimental rating scores, *e.g.*, a 40 point raw score difference on the test gives less than 8 points difference on the rating score.

Another method for evaluation, which is not complicated by the weighting procedure, is to consider the "standard error of estimate", which, for the data analyzed, is computed to be 92.63. Use of this statistic is demonstrated as follows: while a raw test score of 70 through the regression formula predicts an experimental rating of 413, one can through use of the standard error of estimate determine, at rho < .05, the experimental rating actually to lie within the range of 231 to 595. Similarly, the predicted experimental rating, at rho < .05, from a raw test score of 40 is found to be in the range of 183 to 547. Obviously, there is a potential overlap, where persons with raw scores of 40 and 70 on the test may nevertheless obtain the same experimental rating score. It is, furthermore, possible to determine how much of a difference in test scores is required for one to be able to predict, at rho < .05, that the higher-scoring applicant will receive an experimental rating which also is

higher[34]–and this calculation results in a finding that a difference in test scores of over 86 raw points is necessary for such a conclusion to be reached. It should be noted that the total range of raw test scores to this date used to rank successful applicants (*i.e.*, from a low raw score of 48 to the perfect score of 120, which has not been obtained by any) is yet too limited to enable one to say, at the .05 level, that the highest-scoring applicant would be predicted to obtain a higher experimental rating than the lowest-scoring applicant.

***17** Since the 10-C is utilized not only in an attempt to rank the successful candidates, but also to screen the unsuccessful, it is appropriate to analyze the study results with respect to minimum experimental ratings and to predictions for persons scoring at, and below, the test cut-off scores. A test score of 48, the lowest used as a cutoff, yields a predicted experimental rating of 378, or an average unweighted rating on each component of the experimental rating of 3.78 (3 = adequate, 4 = good).[35] Common regression lines can, of course, also be computed separately for each of the twelve components of the rating. When this is done, one finds that a test score of 48 will as to each component predict an unweighted rating of 3 or above, *i.e.*, at least "adequate." While recognizing the risk in extrapolation beyond the range of sample scores,[36] but having little else available for comparable analysis, one can look to estimates of experimental ratings predicted by the regression lines for test scores below 48. If this is done, it appears that even with a test score of 0 the predicted rating is "adequate" or above for the rating as a whole and for seven of the twelve components. Even as to the five components for which the estimated unweighted experimental rating from a 0 test score would be less than 3, the predicted rating cannot be said to be less than "adequate" at rho < .05.

A technique for evaluating tests which employ cut-off scores for screening purposes is to consider "false positives" (persons scoring below the cutoff but nevertheless scoring above the acceptable level of performance on the criterion) in relation to "false acceptances" (persons scoring above the cutoff but below the acceptable level of performance), thereby leading to a comparison between the relative percentage of successful employees above and below the cut-off scores. However, neither this method nor the Taylor-Russell tables (used to estimate net gain in selection accuracy through test usage) can be directly used in the present case because all employees for whom "success" data are available have been screened by the test. It is possible to project the "base rate" (through use of the regression line, standard error of estimate, and normal distribution curves) and then to conduct such inquiries; and, if this be done, one finds any incremental validity to be negligible.

Still another approach is to estimate the effect of the test not on the percentage of persons exceeding minimum performance, but on overall performance of the selected persons. A table given by Anastasi, *op cit.* at p. 173, gives the expected rise in criterion scores through test usage in relation to its validity coefficient and the selection ratio. In the present case, with a .21 coefficient and a selection rate of .19 (506 officers hired from 2721 applicants), and with a standard deviation of known criterion scores of 94.74, one finds from the table that use of the test (had the applicants actually been hired in the order of the test scores) would probably have produced an average gain of 27 points in the total weighted experimental rating (over the rating expected had the test not been used). This gain is equivalent to being rated one point higher on three of the 12 rating components. The average unweighted rating on each of the components would, without the test, have been 4.07 (4 = good), which compares to 4.35, using the test.

***18** The assessment of utility of a test which, like the 10-C, has a statistically significant validity, albeit of very low magnitude, must include certain value judgments. One of these involves consideration of the nature of the job in question and the consequences of a faulty hiring decision. There can be little dispute that police officers perform a vital, and sensitive, function in our society. The desirability of "upgrading" of law enforcement has been emphasized in two reports received in evidence, the 1967 Task Force Report on the Police, issued by the President's Commission on Law Enforcement and the Administration of Justice, and the 1973 Report on Police, issued by the National Advisory Commission on Criminal Justice Standards and Goals. Economic costs are also involved, particularly in view of the cost of academy training of officers and the restraints placed upon discharge of marginal officers under the civil service laws.

Without demeaning the importance of law enforcement officials, however, it can hardly be said that the possibility of occasional selection of an inept officer presents the same type of daily economic and human risk factors as is involved, for example, in the employment of airline pilots or bus-drivers. *Cf. Spurlock v. United Airlines, Inc.*, [5 EPD P 7996] 475 F.2d 216, 219 (CA7 1972); *Usery v. Tamiami Trail Tours, Inc.*, [11 EPD P 10,916] 531 F.2d 224 (CA5 1976) (Age Discrimination in Employment Act case, with somewhat related question). A twelve months' probationary period is provided, during which time the occasional incompetent may be detected and dismissed and during part of which time the employee is undergoing training rather than being "on the street". The principal public concern, it would appear, is not so much that the most able officers be employed (though that, certainly, would be desirable) as that the emotionally unfit not be employed. In this context, it is perhaps noteworthy that the 1967 Presidential Commission's report contained a recommendation for use of psychological tests to detect applicants with personality defects (but no such recommendation respecting aptitude tests); and the 1973 National Advisory Commission's report, while acknowledging the desirability of valid aptitude tests, was skeptical

as to the results of research to that date. So far as the court has been informed, the 10-C was not designed, and has not been validated, for use in detecting emotional disorders or defects.

The DOJ Guidelines § 12b(b), provide that, in determining operational appropriateness, one should consider "the degree of adverse impact of the procedure, the availability of other selection procedures of greater or substantially equal validity, and the need of an employer, required by law or regulation to follow merit principles, to have an objective system of selection." Obviously this latter factor (requirement under civil service law to give some objective test) cannot by itself suffice as justification for a test which has, as here, substantial adverse impact on a racial group. While it can be said that no other available[37] selection with greater validity than the 10-C has been found, yet it must also be said–considering the minimal benefits resulting from the 10-C in the context of this employment situation–that no-test-at-all has "substantially" the same practical validity as the 10-C.

***19** In summary, the 20-B Firefighter test has not been shown to be a valid predictor of a job-relevant criterion measure and the 10-B Policeman test, while having a statistically significant relationship of a very low magnitude with a job-relevant criterion measure, has not been shown to be appropriate for operational use in screening or ranking applicants.[38]

Violation and Remedy

Having concluded that use of the 10-C and 20-B has had an adverse impact upon black applicants and that the studies presented fail to demonstrate job-relatedness, the court must nevertheless determine when the requirements of law were violated and what relief is appropriate therefor. This inquiry should involve no less care than consideration of the tests themselves.

The requirements of Title VII first became applicable to the Personnel Board in March 1972. At that time, and for many years earlier, the Board was required by state law to administer appropriate tests to screen and rank applicants–a requirement which continues to the present time, subject to any over-riding proscriptions of Title VII. It had several years earlier selected the 10-C and 20-B tests as the best tests then available, with the hope that black applicants would fare better than under previous tests. By March 1972 a preliminary, in-house validity study had been conducted, which reflected some improvement in hiring of blacks and the indication of appropriate validity based upon relationship with existing criterion measures. An in-depth independent validation study was immediately undertaken, including investigation of alternative or supplemental selection procedures to improve the predictive validity or decrease adverse impact upon blacks. At least since 1965 the Board has not intentionally discriminated against black applicants but, to the contrary, has attempted to increase black employment within the options available under state law, including modification of the scoring key for the 10-C when recommended by the consultants as a method for increasing validity of the test for black applicants.

The preliminary reports from the consultants, made while more trustworthy measures of job performance were being developed, contained signs of potential validity and recommended continued usage of the test pending the additional studies. Not until April 25, 1975, with respect to the 10-C, and July 8, 1976, with respect to the 20-B, were the studies using these new criterion measures completed and reported to the Board. It was on these respective dates that, in the court's opinion, it should have been concluded that provisional use of the tests was no longer permissible. Prior thereto, the Board was, in the court's opinion, justified in continuing to use the tests (and the eligibility lists generated therefrom) in anticipation of favorable results from those studies. Use of the tests (or of the eligibility lists therefrom) was thereafter, however, contrary to the requirements of Title VII, which override state law inconsistent therewith.

***20** The remedy should be appropriate to the violation found. In this case, from X-11, it is found that, for the two administrations of the 10-C from which eligibility lists used after April 25, 1975, were formed, 658 (or 88%) of the 747 white applicants were placed on the eligibility lists. Had a like percentage of black applicants been so placed, a total of 252 would have been on the lists–128 more than actually placed on the list. Accordingly, to the extent they are still interested, an additional 128 blacks from the prior administrations of the test should be added to the present eligibility lists. This remedy only relates to prohibited use of the 10-C as a screening instrument. An additional measure is needed to correct for the improper use of the test as ranking procedure. Had the eligibility lists been representative of the applicant group and had certifications from the list likewise been representative of the racial composition of the list, approximately 28% of the persons certified would have been black. It is clear that there has been "under-certification" of blacks by this standard, although the precise degree cannot be determined from evidence before the court, which gives such information only by calendar years. The Board is directed to ascertain the extent of such under-certification and in future certifications to include

at least 1 black candidate for every 3 certified until such time that, considering the certifications after April 25, 1975, total number of blacks certified becomes 28% of the total number of persons certified. Thereafter (and until some new selection procedures are adopted which are sufficiently job-related or which have no adverse impact upon blacks) at least 2 of every 7 persons certified by the Board from the revised present list shall be black, provided there be a sufficient number of black applicants interested.

A similar investigation of X-11 with respect to the 20-B, where only one eligibility list has been in effect since July 8, 1976 (the date of the report involving the experimental ratings), results in a conclusion that 91 black applicants should be added to the present eligibility list for firefighters, that at least 1 of every 3 persons hereafter certified shall be black until such time that (considering certifications after July 8, 1976) the total number of blacks certified becomes 14%[39] of the total number of persons certified, and that thereafter (pending adoption of some other valid or non-discriminatory selection instrument) at least 1 of every 7 certified by the Board from the revised current list shall be black.

This order does not preclude use of the 10-C or 20-B as a device for ranking one white as against another white, or one black as against another black. Such a use may be made by the Board, if it so desires, without any discriminatory impact on a racial group. The order does not prevent the Board from new administrations of the 10-C or 20-B (or other tests) or from forming new eligibility lists from time to time; provided, however, that, unless and until a selection instrument is found which either has no adverse impact racially or is sufficiently valid, the test results shall be used in a manner consistent with this opinion, *i.e.*, the eligibility list and certifications to be representative racially of the applicant group regardless of test scores.

Order

***21** Pursuant to the findings and conclusions contained in the Memorandum of Opinion filed herewith, unavoidably protracted because of the need to detail the findings of fact and the reasons therefor, it is ordered as follows:

1. Use by the Personnel Board of Jefferson County of the 30-B Office Workers test has not violated Title VII or other applicable law.

2. Use by the Personnel Board of Jefferson County of the 10-C Policeman Test and the 20-B Firefighter Test has violated Title VII since April 25, 1975, and July 8, 1976, respectively.

3. To the current eligibility list for police officers and deputy sheriffs the Personnel Board shall add the names of 128 black applicants from prior administrations of the 10-C to the extent such number are still interested. In future certifications, at least 1 black on the revised eligibility list shall be certified for each 3 persons certified until such time that the total number of blacks certified after April 25, 1975, shall be 28% of the total number so certified. Thereafter during use of the current eligibility list as so revised, at least two persons of every seven certified shall be black. Pending adoption of some selection procedure which either has no adverse effect upon black applicants or is sufficiently job-related, the number of blacks on any new eligibility list (and certified therefrom) shall be representative of the number of the black applicants.

4. To the current eligibility list for firefighters, the Personnel Board shall add the names of 91 black applicants from prior administrations of the 20-B to the extent such number are still interested. In future certifications, at least 1 black on the revised eligibility list shall be certified for each 3 persons certified until such time that the total number of blacks certified after July 8, 1976, shall be 14% of the total number so certified. Thereafter, during use of the current eligibility list as so revised, at least one person of every seven certified shall be black. Pending adoption of some selection procedure which either has no adverse impact upon black applicants or is sufficiently job-related, the number of blacks on any new eligibility list (and certified therefrom) shall be representative of the number of black applicants.

5. In accordance with F.R.Civ.P. Rule 55(b), the court expressly determines that there is no just reason for delay and expressly directs entry of judgment as to the issues here involved, namely, whether use by the Personnel Board of the 10-C, 20-B, and 30-B tests are proscribed by law and, if so, the appropriate remedy therefor.

Parallel Citations

14 Fair Empl.Prac.Cas. (BNA) 670, 13 Empl. Prac. Dec. P 11,504

Footnotes

1 Fourteen separate county and municipal employers are covered by the law. Cities with a population of under 5,000 are excluded.

2 With multiple vacancies, the number of persons certified is two more than the number of vacancies to be filled.

3 Not presently at issue are requirements (such as age or education) which may be imposed as conditions to taking an examination, nor are specifications (such as residence within Jefferson County) which may give preference to certain applicants.

4 Unless otherwise noted, reference to police officers in the balance of this opinion will also refer to deputy sheriffs.

5 Under F.R.Civ.P. Rule 42, the four actions were consolidated with respect to challenges to Personnel Board tests and a separate trial was scheduled respecting the attacks on the Policeman 10-C, Firefighter 20-B, and Office Worker 30-B tests. At the trial the plaintiffs indicated that the attack on the Office Worker 30-B test was dropped for lack of evidence of adverse impact and that any attack on the 10-C and 20-B tests based on sex was likewise dropped for lack of evidence.

6 The 10-C and 20-B tests were adopted by the Board after initially experimenting, commencing in January 1966, with alternate forms of the IPMA tests.

7 The Board's studies resulted in selection of the 10-C form because of its significant and positive correlation with a greater number of the selected criteria measures than did the alternate IPMA form. The study indicated a significant and positive correlation between 10-C scores and training academy average (as well as several course grades in the academy) and between the training academy average and the officers' latest efficiency ratings.

8 Not until 1968 was the residency requirement of the City of Birmingham removed by the city ordinance. A Jefferson County preference remains in effect, but this can hardly disadvantage blacks, who constitute a larger portion of the Jefferson County population than of neighboring counties.

9 An intent to discriminate would presumably be required for there to be a violation of 42 U.S.C. § 1983, if not of 42 U.S.C. § 1981. See *Washington v. Davis*,.– U.S. – (June 7, 1976).

10 *Albermarle Paper Co. v. Moody*, [9 EPD P 10,230] 422 U.S. 405, 425 (1975).

11 *Griggs v. Duke Power Co.*, [3 EPD P 8137] 401 U.S. 424, 432 (1971).

12 There are some conflicts between the EEOC and the DOJ Guidelines. However, it is not necessary as to the issues presently before the court that a choice be made between the two.

13 Information for this table has been taken from X-11 and from data respecting hires supplied by the parties at the court's request following formal close of the evidence. Certain caveats should be noted: The results of the 10-C exam administered on April 29-30, 1971, have been eliminated because it was not used for employment decisions after March 24, 1972. The results of the 10-C exam administered on September 30 and October 1, 1971 and of the 20-B exam administered on May 26-27, 1971, have been included in the tabulation, even though in part the eligibility lists taken therefrom would have been used prior to March 24, 1972. The number of hires includes those hired in 1972 prior to March 24, 1972. As the Board points out, the number of hires is affected by voluntary choices of the candidates (such as declining job offers or waiving consideration), but, lacking reliable data on such matters for both whites and blacks, the court has looked to actual hires as the measure of the overall selection ratios. Finally, it should be noted that, since persons are permitted to take exams more than once, the applicant figures do not completely accurately reflect the number of different individuals involved. These limitations do not, in the court's opinion, prevent meaningful usage of the data for the purposes indicated.

14 These figures are derived from X-12 and are subject to the appropriate caveats indicated in fn. 13, *supra*. Moreover, X-12 does not have any information for one eligibility list and does not contain percentile information as to several lists. In a very real sense, "passing" an exam is measured not by obtaining a derived score of at least 70 (and thereby being entered on the eligibility list), but by obtaining a score sufficiently high to be placed on the eligibility list at a position where, during use of that list, the candidate will actually be certified to an employing agency.

15 Until a couple of months prior to trial, the litigation was being prepared with the anticipation that all tests under challenge were to be considered at a single hearing. When the decision was made by the court that the first trial would only concern the 10-C, 20-B, and 30-B tests, counsel and witnesses were freed to shift their attention to "loose-ends" on these three tests.

16 Some items on each test elicit knowledge which apparently would be needed for performance of job functions; others do not. Some effort has been made by the test developer to give "face validity", as by expressing an item involving numerical problem solving in the context of information with which job occupants would be dealing. Face validity does not affect validity for usage as a selection procedure so much as it may overcome motivational resistance by those taking the test.

17 "Reliability refers to the consistency of scores obtained by the same persons when reexamined on the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions." Anastasi, PSYCHOLOGICAL TESTING, p. 103 (4th Ed. 1976). The methods used in this study to estimate reliability were the split-half technique (corrected by the Spearman-Brown prophecy formula) and the Kuder-Richardson formula 20.

18 A test item correctly answered by too high a proportion of the applicants is considered not difficult enough; correctly answered by too low a proportion, it is considered too difficult. In this study items correctly answered by 30 to 70% of the applicants were considered satisfactorily difficult.

19 Item "discrimination" is in essence a comparison between scores made on an individual test item and scores made on the total test, thereby ascertaining whether particular items "discriminate" significantly in predicting success on the test as a whole. In the study, significance was established at rho < .05.

20 As previously indicated, the 10-C exam was in fact modified through use, effective April 10, 1974, of an 80-item answer key, which had the effect of eliminating for scoring purposes 40 of the items. This particular change was done to improve correlation with an academy average criterion for blacks (rather than to improve item difficulty or discrimination levels), but it indicates the search for increased test utility. The consultants also recommended consideration of possible modification (or supplementation) of 20-B to improve levels of item difficulty and discrimination. Only one administration of the 20-B was given after this recommendation, and that was done when the existing eligibility list had almost been exhausted but mass administration of an additional test (the PAS) had not yet become feasible.

21 Neither the EEOC nor the DOJ Guidelines require reporting of raw data statistics for SIGMAx, $SIGMAx^2$, SIGMAy, $SIGMAy^2$, or SIGMAxy. Such information would, however, be helpful, permitting application of statistical measures not chosen by authors of the report without the loss of accuracy which results from derivation of such items from means, standard deviations and correlation coefficients.

22 The study of the sample of 59 police officers for whom Raven and PAS test scores were available is subject to question for voluntarism. See A.P.A. Standard E6.1.2.

23 The possibility of bias is of particular concern where subjective evaluations are used as criteria and there are significant differences in those measures for different racial groups. See DOJ § 12b(2). In each study where means and standard deviations are presented separately for blacks and whites with respect to one or more of the basic criterion-measures (academy average, efficiency rating, or experimental rating), the means for blacks is less than for whites, and only in the study which involved but 8 blacks were any of these differences not statistically significant at least at rho < .05. It may be argued that the academy grades should be treated as objective (being based in major part upon multiple-choice exams), but a substantial part of those grades is apparently dependent upon instructors' subjective appraisal of students' performance. (This latter comment is not intended to suggest that more paper-and-pencil tests should be used in the academies, but rather that even academy grades are at least in part subjective measures.)

24 See fn. 17, *supra*. Inquiry into the reliability of a criterion is not a trivial consideration. See A.P.A. Standards E4.4. That the various techniques for estimating such reliability have their own limitations affects the interpretation to be reached, not the desirability of making the effort.

25 Of course, relative standing or grades in the academy may, depending upon the statistical technique employed, be of use in correlating test scores with successful completion of the training.

26 The DOJ Guidelines, § 12b(7)(v)(1), do not require analysis where less than thirty persons are in either of the subgroups.

27 Scattergrams in the present case do indicate, within the ranges of scores available, that relationships between the predictors and criteria for both racial groups are essentially linear.

28 One should view with caution differential studies where either group has less than 30 members. However, the principal concern is that true differences will not appear to be significant with smaller sample numbers.

29 The court has analyzed the 96-subject sample rather than the 189-subject sample because the report of the former includes deviation and correlation data not only for the two racial groups but also for the sample as a whole, facilitating analysis. It should be recognized that analysis from such statistics are subject to rounding errors which could have been avoided had raw data summations been provided.

30 Absent "adverse impact" on whites as a whole, the "underpredicted" whites cannot challenge the test under Title VII.

31 Regression lines should not generally be used for prediction based on predictor scores which are beyond the range of predictor scores found in the sample, for beyond such known scores the relationship may cease to be significant, may cease to be linear, or may have a slope change. Of course the same possibilities exist when one attempts to justify non-selection based upon correlation coefficients developed through the subjects selected or when one attempts to modify coefficients for restriction of range, such as Drs. Farrar and McLaurin have done.

32 The correlations between the 10-C and efficiency ratings were significant only with respect to the volunteer group of 59 officers.

33 The Anastasi volume was qualified during trial as a recognized treatise under F.R.E. 803(18). Additional standard texts used by the court for the purpose of taking judicial notice of basic statistical formulae are Guilford & Fruchter, Fundamental Statistics in Psychology and Education (5th Ed. 1973); Walker & Lev, Statistical Ineference (1953); Burrington & May, Handbook of Probability and Statistics (2nd Ed. 1970); and Mehrens & Lehmann, Standardized Tests in Education (2nd Ed. 1975).

34 A note should be made of this technique since not directly given in most texts [sic]. Using the normal curve, the possibility of scores exceeding $^{??}/_{sigma}$ = + .76 is .2236 and likewise the possibility of scores being less than $^{??}/_{sigma}$ = - .76 is .2236. The possibility of both events occurring is $.2236^2$, or .05. Accordingly, there must be a separation of 2(.76) X standard error of estimate for two predicted scores to be different at rho < .05. One can then determine the difference in predictor scores necessary to produce this separation in predicted scores.

35 By comparison, a raw-test score of 106 (the highest reported in the study for any subject) yields a predicted experimental rating of 472, or an average unweighted rating on each component of 4.73 (5 = very good).

36 See fn. 31, *supra*, and the discussion on page 21 of this opinion respecting assumption of normal distribution when a correlation coefficient is corrected for restriction of range.

37 Studies of the PAS, developed by Drs. Farrar and McLaurin, show extremely high correlations with experimental ratings, as well as with academy training and efficiency ratings. Ironically, the results are so promising as to cause some concern as to a spurious relationship which may not be replicated. In any event, technical difficulties, unresolved to date, have prevented its administration on a wide-scale basis such as for all applicants, so that for practical purposes it is not "available".

38 Correlation studies respecting efficiency ratings and "experimental" ratings were conducted only for officers employed by the City of Birmingham. However, the evidence is persuasive that job requirements for deputy sheriffs and for police officers employed by other municipalities are essentially the same as for Birmingham officers. The higher correlations found with respect to academy training are not of themselves sufficient to justify a conclusion as to operational validity for these officers different from that reached respecting Birmingham.

39 Blacks constituted only 14% of the applicants on the only administration of the 20-B involved.